

Joint Intermodal and Intramodal Label Transfers for Extremely Rare or Unseen Classes

Guo-Jun Qi, *Member, IEEE*, Wei Liu, Charu Aggarwal, *Fellow, IEEE*,
and Thomas Huang, *Life Fellow, IEEE*

Abstract—In this paper, we present a label transfer model from texts to images for image classification tasks. The problem of image classification is often much more challenging than text classification. On one hand, labeled text data is more widely available than the labeled images for classification tasks. On the other hand, text data tends to have natural semantic interpretability, and they are often more directly related to class labels. On the contrary, the image features are not directly related to concepts inherent in class labels. One of our goals in this paper is to develop a model for revealing the functional relationships between text and image features as to *directly transfer intermodal and intramodal labels* to annotate the images. This is implemented by learning a transfer function as a bridge to propagate the labels between two multimodal spaces. However, the intermodal label transfers could be undermined by blindly transferring the labels of noisy texts to annotate images. To mitigate this problem, we present an intramodal label transfer process, which complements the intermodal label transfer by transferring the image labels instead when relevant text is absent from the source corpus. In addition, we generalize the inter-modal label transfer to zero-shot learning scenario where there are only text examples available to label unseen classes of images without any positive image examples. We evaluate our algorithm on an image classification task and show the effectiveness with respect to the other compared algorithms.

Index Terms—Multimodal analysis, intermodal and intramodal label transfers (I2LT), image classification, zero-shot learning

1 INTRODUCTION

Label transfer between different modalities is a problem of using the training examples in one modality (e.g., texts) to enhance the training process for another modality (e.g., images) [14] [35]. This has been studied before in different tasks involving the data of diverse modalities. One of the most important applications is content-based search and semantic indexing for text documents and images. The text documents are much easier to label as compared to associated images on a webpage. Also, since classifiers naturally work better with features that have semantic interpretability, text features are inherently friendly to the classification process in a way that is often a challenge for visual representations of images. It makes it easier to interpret and solve the classification problem in the text modality, while there is a tremendous semantic gap between visual features and the concepts for images. In addition, the challenges of image classification are particularly evident, when the amount of training data available is limited. In such cases, the image classification is further hampered by the paucity of labels.

In the case of images, it is desirable to obtain a feature representation which relates more directly to semantic concepts; a process which will improve the quality of classification. Furthermore, this often has to be achieved with the use of only a limited amount of labeled image

data. This naturally motivates an approach for utilizing the labeled data in the text modality in order to improve image classification. Hence, we implemented an *intermodal label transfer* process in which a transfer function is built to reveal the alignment between modalities so the labels can be transferred across different modalities [35]. We showed that the transfer of the rich label information from texts to images provides much more effective learning algorithms.

Although intermodal label transfer has shown promising result [14] [35], however, we have observed that it might fail when the labels cannot be well aligned between modalities. For example, the text labels of “building” may refer to a large variety of building architectures in different documents, while a test image of “building” often has a certain style of appearance. It is risky to blindly transfer text labels no matter when the visual appearance does not match with any text descriptions from a source corpus. This causes the “negative transfer” problem that refers to transferring of irrelevant information between different modalities [42]. To prevent the negative transfer, we will present an intramodal label transfer process to complement the intermodal label transfer, which will take over the annotation of a test image by transferring image labels in absence of labeled relevant text documents. As a result, this yields a joint Intermodal and Intramodal Label Transfer (I2LT) algorithm, which combines the advantages of both image labels and text labels in the context of a label transfer task.

Formally, we seek to develop a label transfer algorithm for jointly sharing labels across and within different modalities [47] [39] [14]. Specifically, it is applied to the image classification problem in order to leverage the labels in text corpora to annotate image corpora with scarce labels. Such algorithms typically transfer labeling information between

- G.-J. Qi is with the Department of Computer Science, University of Central Florida, Orlando, FL, 32816.
E-mail: guojun.qi@ucf.edu
- W. Liu and C. Aggarwal are with the IBM T.J. Watson Research Center, Yorktown Heights, NY 10598.
E-mail: {weiliu, charu}@us.ibm.com
- T. Huang is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL, 61801.
E-mail: huang@ifp.uiuc.edu

heterogeneous feature spaces [14] [50] [48] instead of homogeneous feature spaces [47]. Heterogeneous transfer learning is usually much more challenging due to the unknown alignment across the distinct feature spaces. In order to bridge across two distinct feature spaces, the key ingredient is a “transfer function” which can explain the alignment between text and image feature spaces through the use of a feature transformation. This transformation is used for the purpose of effective image classification and semantic indexing. As discussed earlier, it is achieved with the use of co-occurrence data that is often available in many practical settings. For example, in many real web and social media applications, it is possible to obtain many *co-occurrence pairs* between text and images [36]; in web pages, the images are surrounded by text descriptions on the same web page. Similarly, there is a tremendous amount of linkage between text and images on the web, through comments in image sharing sites, posts in a social networks, and other linked text and image corpora. It is reasonable to assume that the content of the text and the images are highly correlated in both scenarios. This information provides a *semantic bridge*, which can be exploited in order to learn the alignment and label transfer between the different modalities.

In contrast to previous work [14] [50] [48], the label transfer proposed in this paper can establish the alignment between texts and images even if **the new test images do not have any surrounding text description**, or if the co-occurrence data is independent of the labeled source texts. This increases the flexibility of the algorithm and makes it more widely applicable in many practical applications. Specifically, in order to perform the label transfer process, we create a new *topic space* into which both the text and images are mapped. Both the occurrence set and training set are used to learn the transfer function, which aligns heterogeneous text and image spaces. We also follow the *principle of parsimony*, and encode as few topics as possible in order to align between text and images for regularization. This principle has a preference for the least complex model, as long as the text and image alignment can be well explained by the learned transfer function. After the transfer function is learned, the labels can be propagated from any labeled text corpus to any new image by intermodal label propagation. While labels from the images are also used for improving accuracy, one characteristic of our transfer function is that it is particularly robust in the presence of a very small number of scarce training examples.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work. Then we propose an intermodal label transfer process in Section 3 and show how the labels of text corpus can be propagated to image corpus. In section 4, a joint Intermodal and Intramodal Label Transfer (I2LT) process is proposed, along with a transfer function in Section 5 that instantiates the joint model. In Section 6, we present the objective problem along with a proximal gradient based algorithm for solving the optimization problem. We also present a zero-shot learning extension of the proposed algorithm to classify images of unseen classes in Section 7. The experiment results are presented in section 8. The conclusion and summary is presented in Section 9.

2 RELATED WORK

A variety of transfer learning methods have been proposed in prior pioneering works, e.g., domain adaption [15], [25], [26], [39], [40], [47], cross-category information sharing [38], and heterogeneous transfer learning [14], [18], [35], [50]. In this paper, we concentrate on learning cross-modal correspondence and sharing the semantic information across different modalities.

Learning semantic correspondence from text to images can be seen as a transfer learning problem that involves heterogeneous data points across different feature spaces. For example, [50] proposes *heterogeneous transfer learning*, which uses both user tags and related document text as auxiliary information to extract a new latent feature representation for each image. However, it does not utilize the text labels to enrich the semantic labels of images, which may restrict its performance when the image labels are very scarce. On the other hand, translated learning [14] attempts to label the target instances through a Markovian chain. A translator is assumed to be available between source and target data for correspondence. However, given an arbitrary new image, such a correspondence is not always directly available between any text and image instances. In this case, a generative model is used in the Markovian chain to construct feature-feature co-occurrence. This model is not reliable when co-occurrence data is noisy and sparse. On the contrary, we explicitly learn a semantic transfer function, which directly propagates semantic labels from text to images even if the semantic correspondence is not available beforehand for a new image. It avoids overfitting into the noisy and sparse co-occurrence data by imposing the prior of fewest topics on semantic translation.

It is also worth noting that learning label transfer across heterogeneous modalities is different from the conventional *heterogeneous learning*, such as multi-kernel learning [4] and co-training [7]. In heterogeneous learning, each instance must contain different views. On the contrary, when translating text to images [35], *it is not required that an image has an associated text view*. This makes the problem much more challenging. The correspondence between text and images is established by the learned transfer function, and a single image view of an input instance is enough to predict its label by a label transfer process.

We also distinguish the proposed label transfer model from the other latent models. Previous latent methods, such as Latent Semantic Analysis [24], Probabilistic Latent Semantic Analysis [20], Latent Dirichlet Allocation [6] and Multimodal Latent Attributes [17], are restricted to latent factor discovery from the co-occurrence observations. On the contrary, in this paper, the goal is to establish semantic bridge so that the discriminative labeling information can be propagated between the source and target spaces. To the best of our knowledge, it is one of the first algorithms to address such heterogeneous label transfer problem *via a parsimonious latent topic space*. It is worth noting that even with *unknown correspondence to source instances*, it can still label the new instance by predicting its correspondence based on the learned transfer function.

We also note that the proposed label transfer problem also differs from the problem of translating images and

videos into sentences of natural languages [45] [33]. Usually Recurrent Neural Networks (RNNs) [19] are used as a mathematical tool to map the visual feature vectors into words via a sequence of intermediate representation of long short-term memory cells [45]. The translation problem has been recognized as a very challenging task, since it requires the machine not only capable of reading the content of images and videos accurately, but also be able to translate the visual elements into sentences in a correct order with a satisfactory level of grammatical correctness. In this paper, we do not aim to solve this challenging problem. Instead we consider the label transfer from texts to images, where we do not need to compose the sentences. Also, our goal differs from composing the description of the visual content in sentences. Instead, we wish to utilize the abundant labeled text documents to improve the classification accuracy for the image classification tasks.

Although we focus on label transfer from texts to images, the model developed in this paper is equally applicable to the other label transfer tasks between different modalities. For example, the previous work has demonstrated an application where the labels of English documents are transferred to annotate the Chinese documents [14]. Similarly, the speech segments can be aligned by learning a transfer function by which the labels can be transferred across different languages to annotate the speeches. The label transfer model can also be applied for audio-video recognition tasks [11], [29]. Similar to the scenario set in this paper, a test audio will have no paralleled video and it must be aligned to the existing corpus of videos to enable intermodal label transfer. But [29] explores a slightly different idea – instead of aligning the test sample with the video corpus, they attempt to reconstruct the paralleled video through multimodal deep networks [30]. This approach is indirect for label transfer and an independent classifier must be trained for audio-video recognition tasks.

In an earlier work [18], Flickr images with tags have been used to learn several CCA variants for cross-modal retrieval task. They incorporated a third view of supervised semantic information or unsupervised word clusters to bridge the cross-modal gap, along with the visual and text views. On the contrary, an important byproduct of the proposed algorithm is a intermodal transfer function, which can measure the cross-modal relevance directly. It is also learned with the supervised labeled image/text pairs. In this spirit, our approach also involves a “third view” of the labeled concepts. However, our approach is motivated to annotate the labels of semantic concepts, rather than learning the cross-modal relevance directly. This makes the proposed approach in a complimentary technical line to those CCA variants presented in [18].

A more recent work [32] proposed to use privileged information to augment Support Vector Machines (SVMs). Additional training bags were collected from textural descriptions of images, where positive bags contain the returned images containing relevant tags, while negative bags do not contain any images with relevant tags. Then the problem with the training bags was formulated as multi-instance learning problem, and positive bags provide privileged information to augment the training of the classifiers with more positive instances of images. Our approach dif-

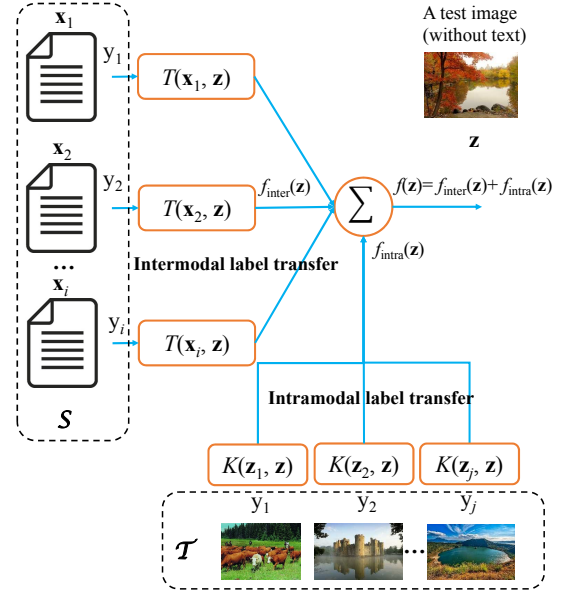


Fig. 1. Illustration of semantic label propagation from text to images by the learned transfer function. On the left is the labeled text corpus S , and at the bottom is the labeled image corpus T . The proposed I2LT transfers the labels from both corpora to annotate a test image at the top right corner. The output label is given by a discriminant function $f(z)$. Note that the test image is not associated with any text document. Hence, the transfer function T is applied for the intermodal label transfer f_{inter} from source corpus S , along with the intramodal label transfer f_{intra} from the image corpus T .

fers from this method in directly transferring the text labels to reconstruct image labels, rather than training an image classifier. However, both methods do not assume the availability of text information for testing images, making them applicable to label new images without text descriptions.

3 INTERMODAL LABEL TRANSFER

In this section, we will introduce the notations and problem definitions for the label transfer process. Let $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Z} \subset \mathbb{R}^q$ be the source and target feature spaces, which have a dimensionality of p and q respectively. For the purpose of this paper, the source space corresponds to the text modality, and the target space corresponds to the image modality. In the source (text) space, we have a set of n text documents in \mathcal{X} . Each text document is represented by a feature vector $\mathbf{x}_i \in \mathcal{X}$. This text corpus, $S = \{(\mathbf{x}_i, y_i) | 1 \leq i \leq n\}$, has already been annotated with class labels, where $y_i \in \{+1, -1\}$ is the binary label for each document \mathbf{x}_i . The binary assumption is made to avoid notational clutter, and it can be straightforwardly extended to encode multiple classes.

The images are represented by feature vectors \mathbf{z} in the target space \mathcal{Z} . The task is to relate the feature structure of the source (text) space to the target space (image) space, so that the labeling information can be shared between two spaces. The goal of the transformation process is to provide a classifier for the target (image) domain in the presence of scarce labeled data for the latter domain.

In order to perform the label propagation from the text to the image domain, we need a bridge, which relates the text and image information. A key component which provides such bridging information about the relationship between the text space \mathcal{X} and image feature space \mathcal{Z} is a set of *co-occurrence pairs* $\mathcal{C} = \{(\mathbf{x}_k, \mathbf{z}_k) | \mathbf{x}_k \in \mathcal{X}, \mathbf{z}_k \in \mathcal{Z}, k = 1, \dots, l\}$. Such co-occurrence information is abundant in the context of web and social network data. In fact, it may often be the case that the co-occurrence information between text and images can be more readily obtained than the class labels in the target (image) domain. For example, in many web collections, the images may co-occur with the surrounding text on the same web page. Similarly, in web and social networks, it is common to have implicit and explicit links between text and images. Such links can be viewed more generally as co-occurrence data. This co-occurrence set provides the semantic bridge needed for transfer learning.

Besides the co-occurrence set, we also have a small set $\mathcal{T} = \{(\mathbf{z}_j, y_j) | \mathbf{z}_j \in \mathcal{Z}, 1 \leq j \leq m\}$ of labeled images. This is an auxiliary set of training examples, and its size is usually much smaller than that of the set of labeled source examples. In other words, we have $m \ll n$. As we will see, the auxiliary set is used in order to enhance the accuracy of the transfer learning process.

One of the key intermediate steps during this process is the design of a *transfer function* between text and images. This transfer function serves as a conduit to measure the alignment between text and image features. We will show that such a conduit can be used directly in order to propagate the class labels from text to images. The transfer function T is defined jointly on text space \mathcal{X} and image space \mathcal{Z} as $T : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$. It assigns a real value to one pair of texts and image instances to weigh their alignment. This value can be either positive or negative, representing either positive or negative match. Given a new image \mathbf{z} , its label is determined by an intermodal discriminant function as a linear combination of the class labels in \mathcal{S} weighted by the corresponding transfer functions

$$f_{\text{inter}}(\mathbf{z}) = \sum_{i=1}^n y_i T(\mathbf{x}_i, \mathbf{z}) \quad (1)$$

Then, the sign of $f_{\text{inter}}(\mathbf{z})$ decides the class label of \mathbf{z} .

4 JOINT INTERMODAL AND INTRAMODAL LABEL TRANSFERS

In addition to the above inter-modal label transfer model, we can transfer the image labels from the training set \mathcal{T} directly to annotate a test image \mathbf{z} . Formally, we can define the following discriminant function for the intra-modal label transfer:

$$f_{\text{intra}}(\mathbf{z}) = \sum_{j=1}^m y_j \alpha_j K(\mathbf{z}_j, \mathbf{z}) \quad (2)$$

where $\alpha = [\alpha_1, \dots, \alpha_m]'$ are the real-valued coefficients for intra-modal label transfer, and $K(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a kernel function between two images satisfying Mercer's condition (e.g., Gaussian kernel) [13]. This label transfer has the similar form as the discriminant function of kernelized

support vector machine [13], with each nonzero α_j corresponding to a support vector.

Combining the inter-modal label transfer (1) and the intra-modal label transfer (2), we obtain the following discriminant function to decide the label of a test image \mathbf{z} :

$$f(\mathbf{z}) = f_{\text{inter}}(\mathbf{z}) + f_{\text{intra}}(\mathbf{z}) \quad (3)$$

It is worth noting that usually **no** surrounding text document comes with the test image \mathbf{z} . But we can always apply the transfer function to align the test image with the text documents from the source corpus \mathcal{S} . This solves the out-of-sample problem so the text labels can be transferred to annotate any new images.

This extends the inter-modal label transfer paradigm. We expect the intermodal and intramodal label transfers can collaboratively annotate the test images, aggregating both the label information from both texts and images. This can mitigate the negative transfer problem [38] [42] when the text documents in the source corpus cannot properly specify the visual aspect of a test image. In this case, we expect the image labels would take over to annotate the image based on its visual appearance. In this spirit, the intramodal label transfer component plays a role of “watchdog” to overlook and complement the intermodal label transfer. As to be shown in the experiment, it successfully improves the intermodal label transfer model and outperforms the compared algorithms over all the categories for a image classification task.

The learning problem of establishing joint label transfers boils down to learn the coefficients α , along with the transfer function that properly explains the alignment between text and image spaces. This overall process is illustrated intuitively in Figure 1. Since the key to an effective transfer learning process is to learn the function T , we need to formulate an optimization problem which maximizes the classification accuracy obtained from this transfer process. First, we will first set up the optimization problem more generally without assuming any canonical form for T . Later, we will set up a *canonical form* for the transfer function in the form of matrices which represent topic spaces. The parameters of this canonical form will be optimized in order to learn the transfer function. We propose to optimize the following problem to jointly learn the parameters of intermodal and intramodal functions:

$$\begin{aligned} \min_{\alpha_j, T} & \gamma \sum_{j=1}^m \ell(y_j f(\mathbf{z}_j)) + \lambda \sum_{k=1}^l \delta(\mathbf{x}_k, \mathbf{z}_k) + \Omega(T) \\ \text{s.t. } & 0 \leq \alpha_j \leq C, j = 1, \dots, m \end{aligned} \quad (4)$$

where (1) $\ell(\cdot)$ is the loss function of the training errors on the labeled image set \mathcal{T} ; (2) $\delta(\cdot)$ is the loss function that measures the misalignment between the co-occurrence text-image pairs, and minimizing this loss would maximize the value of transfer function over the co-occurrence pairs; and (3) The last term $\Omega(T)$ regularizes the learning of the transfer function in order to improve the generalization performance. In the following section, we will present the detailed forms of these loss functions and the regularizer.

In addition, γ and λ are positive balancing parameters, which define the relative importance of training data and co-occurrence pairs in the objective function; and the bound constraint follows the conventional regularization constraint

on the coefficients in support vector machines [13], which is expected to yield better generalization performance.

Remark on the three data sets $\mathcal{S}, \mathcal{C}, \mathcal{T}$: It is worth noting that the labelled text set \mathcal{S} is used to propagate their labels to annotate the target images. We do not need to set the text part of the co-occurrence set \mathcal{C} to be the same as \mathcal{S} , since the modeling of co-occurrence and the label propagation are different. The labeled image set \mathcal{T} can also differ from \mathcal{C} . These labeled images are used to minimize the classification errors involved in the first term of objective function (4), which is different from maximization of co-occurrence consistency in the second term.

5 INTERMODAL TRANSFER FUNCTION

In this section, we will design the canonical form of the transfer function $T(\mathbf{x}, \mathbf{z})$ in terms of underlying *topic spaces*. This provides a closed form to our transfer function, which can be effectively optimized. Topic spaces provide a natural intermediate representation which can semantically link the information between the text and images. One of the challenges to this is that text and images have inherently different structure to describe their content. For example, text is described in the form of a vector space of sparse words, whereas images are typically defined in the form of feature vectors that encode the visual appearances such as color, texture and their spatial layout. To establish their connection, one must discover a common structure which can be used in order to link them. A text document usually contains several topics which describe different aspects of the underlying concepts at a higher level. For example, in a web page depicting a *bird*, some topics such as the head, body and tail may be described in its textual part. At the same time, there is an accompanying *bird* image illustrating them. By mapping the original text and image feature vectors into a space with several unspecified topics, they can be semantically linked together by investigating their co-occurrence data. By using this idea, we construct two transformation matrices to map text and images into a common (hypothetical) latent topic space with dimension r , as in the previous work [37], which makes them directly comparable. The dimensionality is essentially equal to the number of topics. We note that it is not necessary to know the exact semantics of latent topics. We only attempt to model the semantic correspondence between the unknown topics of text and images. The learning of effective transformation matrices (or, as we will see later, an appropriate function of them) is the key to the success of the semantic translation process. These matrices are defined as follows.

$$\mathbf{W} \in \mathbb{R}^{r \times p} : \mathcal{X} \subset \mathbb{R}^p \rightarrow \mathbb{R}^r, \mathbf{x} \mapsto \mathbf{W}\mathbf{x} \quad (5)$$

$$\mathbf{V} \in \mathbb{R}^{r \times q} : \mathcal{Z} \subset \mathbb{R}^q \rightarrow \mathbb{R}^r, \mathbf{z} \mapsto \mathbf{V}\mathbf{z} \quad (6)$$

The transfer function is defined as a function of the source and target instances by computing the inner product in our hypothetical topic space, with a nonlinear hyperbolic tangent activation $\tanh(\cdot)$

$$\begin{aligned} T(\mathbf{x}, \mathbf{z}) &= \tanh(\langle \mathbf{W}\mathbf{x}, \mathbf{V}\mathbf{z} \rangle) \\ &= \tanh(\mathbf{x}'\mathbf{W}'\mathbf{V}\mathbf{z}) = \tanh(\mathbf{x}'\mathbf{S}\mathbf{z}) \end{aligned} \quad (7)$$

Here $\langle \cdot, \cdot \rangle$ and $'$ denote the inner product and transpose operations respectively. Clearly, the choice of the transformation matrices (or rather the product matrix $\mathbf{W}'\mathbf{V}$) impacts the transfer function T directly. Therefore, we will use the notation \mathbf{S} in order to briefly denote the matrix $\mathbf{W}'\mathbf{V}$. Clearly, it suffices to learn this product matrix \mathbf{S} rather than the two transformation matrices separately. The above definition of the matrix \mathbf{S} can be used to rewrite the inter-modal label transfer function as follows:

$$f_{\text{inter}}(\mathbf{z}) = \sum_{i=1}^n y_i \tanh(\mathbf{x}'_i \mathbf{S} \mathbf{z}) \quad (8)$$

6 OBJECTIVE PROBLEM

Putting together with the intermodal and intramodal label transfer formula in (8) and (2), we define the discriminant function $f(\mathbf{z}) = f_{\text{inter}}(\mathbf{z}) + f_{\text{intra}}(\mathbf{z})$ which can be substituted in the objective function of the optimization problem (4) for learning the transfer function. In addition, we use the conventional squared norm to regularize the transfer function T on two transformations respectively:

$$\Omega(T) = \frac{1}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{V}\|_F^2)$$

Here, the expression $\|\cdot\|_F$ represents the Frobenius norm. Then, we can use the aforementioned substitutions in order to rewrite the objective function of Eq. (4) as follows:

$$\begin{aligned} \min_{\mathbf{S}=\mathbf{W}'\mathbf{V}, \alpha} \gamma \sum_{j=1}^m \ell(y_j f(\mathbf{z}_j)) + \lambda \sum_{k=1}^l \delta(\mathbf{x}'_k \mathbf{S} \mathbf{z}_k) \\ + \frac{1}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{V}\|_F^2) \\ \text{s.t. } 0 \leq \alpha_j \leq C, j = 1, \dots, m \end{aligned} \quad (9)$$

The goal is to determine the value of \mathbf{S} , which optimizes the objective function in Eq. (9). We note that this objective function is not jointly convex in \mathbf{W} and \mathbf{V} . This implies that the optimum value of \mathbf{S} may be hard to find with the use of straightforward gradient descent techniques, which can easily get stuck in local minima. Fortunately, it is possible to learn \mathbf{S} directly from Eq. (9) by the trace norm as in [43] [3]. It is defined as follows:

$$\|\mathbf{S}\|_{\Sigma} = \inf_{\mathbf{S}=\mathbf{W}'\mathbf{V}} \frac{1}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{V}\|_F^2) \quad (10)$$

The trace norm is a convex function of \mathbf{S} , and can be computed as the sum of its singular values. The trace norm is different from the conventional squared norm for regularization purposes, and is actually a surrogate of matrix rank [10], and minimizing it can limit the dimension r of the topic space. In other words, minimizing the trace norm results in the fewest topics to explain the correspondence between text and images. This implies that concise semantic transfer with fewer topics is more effective than tedious translation on cross-domain correspondence between text and images, as long as the learned transfer function complies with the observations (i.e., the co-occurrence and auxiliary data). This is consistent with the parsimony principle, which states preference for the least complex translation model. A parsimonious choice is also helpful in avoiding overfitting problems which may arise in scenarios where the number of auxiliary training examples are small.

The objective function in Eq. (9) can be rewritten as follows with the use of the trace norm:

$$\min_{\mathbf{S}, 0 \leq \alpha \leq C} \gamma \sum_{j=1}^m \ell(y_j f(\mathbf{z}_j)) + \lambda \sum_{k=1}^l \delta(\mathbf{x}'_k \mathbf{S} \mathbf{z}_k) + \|\mathbf{S}\|_{\Sigma} \quad (11)$$

We note that this objective function has a number of properties, which can be leveraged for optimization purposes. In the next section, we discuss the methodology for optimization of this objective function.

6.1 Joint Optimization Algorithm

In order to optimize the objective function above, we first need to decide which functions are used for $\ell(\cdot)$ and $\delta(\cdot)$ in Eq. (11).

Recall that these functions are used to measure compliance with the observed co-occurrence and the margin of discriminant functions $f(\cdot)$ on the auxiliary data set, respectively. In this case, we use the hinge loss $\ell(\tau) \triangleq (1 - \tau)_+$ for the loss function over the training set, where $(\cdot)_+$ denotes the positive component. We choose the hinge loss here because it has been shown to be more robust to the noisy outliers of training examples. Clearly, minimizing the hinge loss tends to maximize the margin $y_j f(\mathbf{z}_j)$.

On the other hand, in compliance with the use of hyperbolic tangent activation in Eq. (7), we choose $\delta(a_k) \triangleq -\log \frac{1}{2}(1 + \tanh(a_k)) = \log(1 + \exp(-2a_k))$ in the objective function (11) with a_k denoting $\mathbf{x}'_k \mathbf{S} \mathbf{z}_k$. This choice of the loss function essentially uses the logistic loss to measure the misalignment made by the transfer function between a co-occurrence pair of \mathbf{x}_k and \mathbf{z}_k . Minimizing this logistic loss tends to maximize the values of the transfer function over the co-occurrence pairs.

The aforementioned substitutions instantiate the objective function (11) which is nonlinear in \mathbf{S} and α . One possibility for optimizing an objective function of the form represented in Eq. (11) is to use the method of Srebro et al. [43]. The work showed that the dual problem can be optimized by the use of semi-definite programming (SDP) techniques. Although many off-the-self SDP solvers use interior point methods and return a pair of primal and dual optimal solutions [9], they do not scale well with the size of the problem. The work in [3] proposes a gradient based method which replaces the non-differentiable trace norm with a smooth proxy. But the smoothed approximation to $\|\mathbf{S}\|_{\Sigma}$ may not guarantee that the obtained minima still correspond to fewest topics for label transfer.

Alternatively, a proximal gradient method is proposed in [44] to minimize such non-linear objective functions with the use of a trace norm regularizer. We will use such an approach to optimize over \mathbf{S} and α_j in an alternating fashion in this paper. In order to represent the objective function of Eq. (11) more succinctly, first we introduce the optimization over \mathbf{S} , and we define the function $F(\mathbf{S})$ as follows.

$$F(\mathbf{S}) = \gamma \sum_{j=1}^m \ell(y_j f(\mathbf{z}_j)) + \lambda \sum_{k=1}^l \delta(\mathbf{x}'_k \mathbf{S} \mathbf{z}_k) \quad (12)$$

Then, the objective function of Eq. (11) can be rewritten as $F(\mathbf{S}) + \|\mathbf{S}\|_{\Sigma}$. In order to optimize this objective function,

the proximal gradient method quadratically approximates it by Taylor expansion at current value of $\mathbf{S} = \mathbf{S}_{\tau}$ with a proper coefficient L as follows:

$$\begin{aligned} Q(\mathbf{S}, \mathbf{S}_{\tau}) &= F(\mathbf{S}_{\tau}) + \langle \partial F(\mathbf{S}_{\tau}), \mathbf{S} - \mathbf{S}_{\tau} \rangle \\ &\quad + \frac{L}{2} \|\mathbf{S} - \mathbf{S}_{\tau}\|_F^2 + \|\mathbf{S}\|_{\Sigma} \\ &= \frac{L}{2} \|\mathbf{S} - \mathbf{G}_{\tau}\|_F^2 + \|\mathbf{S}\|_{\Sigma} + \text{const} \end{aligned} \quad (13)$$

where $\partial F(\mathbf{S})$ denotes the subgradient of F at \mathbf{S} . Here we use the subgradient because of the non-differentiability of loss function $\ell(\cdot)$ at 0. We can further introduce the notation \mathbf{G}_{τ} in order to organize the above expression:

$$\mathbf{G}_{\tau} = \mathbf{S}_{\tau} - L^{-1} \partial F(\mathbf{S}_{\tau}) \quad (14)$$

The subgradient $\partial F(\mathbf{S}_{\tau})$ can be computed as follows:

$$\begin{aligned} \partial F(\mathbf{S}_{\tau}) &= \gamma \sum_{j=1}^m y_j \cdot \partial \ell(y_j f(\mathbf{z}_j)) \nabla_{\mathbf{S}} f(y_j f(\mathbf{z}_j)) \\ &\quad + \lambda \sum_{k=1}^l \{ \partial \delta(\mathbf{x}'_k \mathbf{S}_{\tau} \mathbf{z}_k) \mathbf{x}_k \mathbf{z}'_k \} \end{aligned} \quad (15)$$

where $\partial \ell$ is the subdifferential of $\ell(\cdot)$, $\nabla_{\mathbf{S}} f$ is the gradient of f to \mathbf{S} , and $\partial \delta$ is the derivative of logistic loss as derived before. Then, the matrix \mathbf{S} can be updated by minimizing $Q(\mathbf{S}, \mathbf{S}_{\tau})$ with fixed \mathbf{S}_{τ} iteratively. This can be solved by singular value thresholding [10] with a closed-form solution (see Line 4 in Algorithm 1).

On the other hand, the optimization over α can be performed by using the gradient projection method [5]. With fixed \mathbf{S} at each iteration, each α_j can be updated as

$$\alpha_j \leftarrow \Pi_{[0, C]}(\alpha_j - \epsilon \partial F(\alpha_j))$$

where ϵ is a positive step size, $\Pi_{[0, C]}$ is the projection onto $[0, C]$, and

$$\partial F(\alpha_j) = \gamma \sum_{j'=1}^m \partial \ell(y_{j'} f(\mathbf{z}_{j'})) K(\mathbf{z}_j, \mathbf{z}_{j'})$$

is the subdifferential of F at α_j .

Algorithm 1 summarizes the proximal gradient based method to optimize the expression in Eq. (11). Note that the intermodal discriminant function $f_{\text{inter}}(\mathbf{z})$ is not convex as a function of \mathbf{S} , and hence the objective function is not convex either. But as long as the step size (i.e., L^{-1}) is properly set, the objective function (11) tends to decrease in each iteration, usually converging to a stationary point (may not be a global optimum) [44]. This is different from our previous work [35], where we adopted a linear transfer function yielding a convex objective problem. The nonlinear function has shown better performance on learning alignment between multiple modalities in literature [30] [28].

7 ZERO-SHOT LABEL TRANSFER FOR UNSEEN CLASSES

The goal of zero-shot learning [16], [23], [34] is to build classifiers to label the unseen classes without any training image examples. However, there can be some positive examples available in text modality. In this section, we show that our cross-modal label transfer model can also be used in this setting.

Algorithm 1 Joint Optimization for Problem (11).

input Co-occurrence set \mathcal{C} , labeled text corpus \mathcal{S} , labeled image dataset \mathcal{T} , and balancing parameters λ and γ .

1 Initialize $\mathbf{S}_\tau \leftarrow 0$ and $\tau \leftarrow 0$.

2 Initialize $\alpha_j \leftarrow 0$ for each j .

repeat

3 Set $\mathbf{G}_\tau = \mathbf{S}_\tau - L^{-1} \partial F(\mathbf{S}_\tau)$.

4 Singular Value Thresholding:

$$\mathbf{S}_{\tau+1} \leftarrow \text{Udiag}(\boldsymbol{\sigma} - L^{-1})_+ \mathbf{V}'$$

where $\text{Udiag}(\boldsymbol{\sigma}) \mathbf{V}'$ gives the SVD of \mathbf{G}_τ .

5 Update $\alpha_j \leftarrow \Pi_{[0, C]}(\alpha_j - \epsilon \partial F(\alpha_j))$ for $j = 1, \dots, m$.

6 $\tau \leftarrow \tau + 1$.

until Convergence or maximum iteration number achieves.

Specifically, suppose that we have n_{sc} seen classes with labeled training images, and our goal is to annotate the images for n_{uc} unseen classes. In addition to the images, we have the labeled text examples for both seen and unseen classes. Then, zero-shot label transfer aims to transfer the text labels to annotate the images of unseen classes. In principle, the same inter-modal label transfer function f_{inter} in Eq. (1) is applicable in labeling the images of unseen classes, since the text labels, of no matter seen or unseen classes, can be transferred to label those images. However, in this case, the intra-modal label transfer term f_{intra} will not be used any more since we cannot get access to the image labels of those unseen classes.

The learning of the inter-modal transfer function does not need to be changed to adapt to the zero-shot learning problem. However, for the sake of fair zero-shot learning scenario, we should exclude the co-occurrence text-image pairs belonging to the unseen classes from the training set. Only co-occurrence pairs of seen classes would be used to model the correlation between the text and image modalities via the inter-modal transfer function f_{inter} . This idea of involving pairs of seen classes has been adopted in literature [16], [22] to learn the inter-modal correlations, which plays the critical role in bridging the gap across multi-modalities.

On the other hand, we note that the labeled image examples of seen classes can still be used in training the model, except that they should be treated as negative examples for the unseen classes. These seen classes provide useful auxiliary information to exclude the regions from the feature space where the unseen classes are unlikely to be present¹. This prior has been explored in [16] to improve the classification accuracy for the unseen classes.

We will demonstrate the experiment result in zero-shot learning scenario in Section 8.4.

8 EXPERIMENT

In this section, we compare the proposed label transfer paradigm with a pure image classification algorithm with a SVM classifier based on pure image features, along with the

1. We assume that different classes are exclusive to each other, i.e., we consider a multi-class problem rather than a multi-label problem. This assumption holds for many image classification problems, such as object and face recognitions.

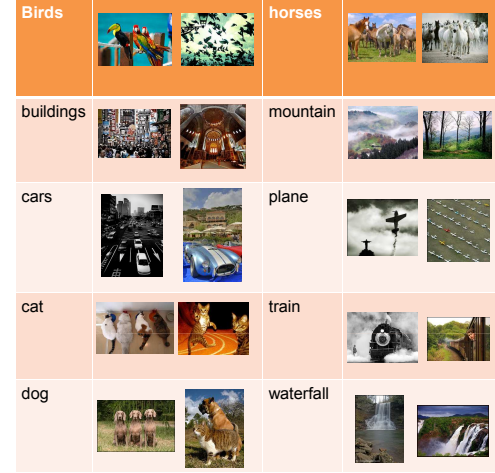


Fig. 2. Examples of images over the different categories.

TABLE 1

The number of occurrence pairs of texts and images for each category.

Category	Occurrence pairs	Category	Occurrence pairs
birds	930	horses	654
buildings	9216	mountain	4153
cars	728	plane	1356
cat	229	train	457
dog	486	waterfall	22006

TABLE 2

The number of positive and negative images for each category.

Category	positive examples	negative examples
birds	338	349
buildings	2301	2388
cars	120	125
cat	67	72
dog	132	142
horses	263	268
mountain	927	1065
plane	509	549
train	52	53
waterfall	5153	5737

TABLE 3

The number of Wiki articles for each category. We collect these articles by retrieving their subcategories.

Category	Number	Category	Number
birds	82	horses	197
buildings	98	mountain	151
cars	146	plane	64
cat	91	train	150
dog	106	waterfall	249

other existing transfer learning methods proposed in [50] [14] [35]. We will show the superior results of our approach to the other methods, with limited amount of training data.

8.1 Setting

We compare the accuracy and sensitivity of our label transfer approach with a number of algorithms below:

1. *SVM* [13]. As the baseline, we directly train the SVM classifiers based on the visual features extracted from

images. This method does not use any of the additional information available in corresponding text in order to improve the effectiveness of target domain classification. The method is also susceptible to the case when we have a small number of test instances.

2. *TLRisk (Translated Learning by minimizing Risk)* [14]. This is another transfer learning algorithm, which performs the translation by minimizing risk (TLRisk) [14]. The algorithm transfers the text labels to image labels via a Markovian chain. It learns a probabilistic model to translate the text labels to image labels by exploring the occurrence relation between text documents and images. We note however, that such an approach does not use the topic-space methodology which is more useful in connecting heterogeneous feature spaces.
3. *HTL (Heterogeneous Transfer Learning)* [50]: This algorithm is the best fit to our scenario with heterogeneous spaces compared to other transfer learning algorithms such as [40] [39] on a homogeneous space. This method has also been reported to achieve superior effectiveness results. It maps each image into a latent vector space where an implicit distance function is formulated. In order to do so, it also makes use of the occurrence information between images and text documents as well as images and visual words. To facilitate this method into our scenario, user tags in *Flickr* are extracted to construct the relational matrix between images and tags as well as that between tags and documents. Images are represented in a new feature space on which the images can be classified by applying the k -nearest neighbor classifier (here k is set to be 3) based on the distances in the new space. We refer to this method as **HTL**.
4. *Translator from Text to Images (TTI)* [35]: This is our previous label transfer algorithm which only uses intermodal label transfer without considering the intramodal label transfer. This model fails to outperform the other compared algorithms on some categories [35]. As aforementioned, this might be caused by the misalignment between text documents and test images.
5. *Joint Intermodal and Intramodal Label Transfer (I2LT)*: this is the proposed approach in this paper.

In the experiments, a small number of training images are randomly selected from each category as labeled instances in \mathcal{T} for the classifiers. The remaining images in each category are used for testing the performance of the classification task. Only a small number of training examples are used, making the problem very challenging from the training perspective. This process is repeated five times. The error rate and the standard deviation for each category is reported in order to evaluate the effectiveness of the compared classifiers. We also use varying number of co-occured text-image pairs to construct the classifier, and compare the corresponding results with related algorithms.

In the experiments, the parameters λ , γ (used to decide the importance of auxiliary data and co-occurrence data from the objective function in (11)) and C (used to

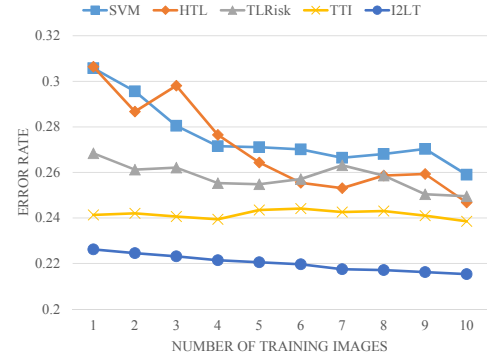


Fig. 3. Average error rate of different algorithms with varying number of training images.

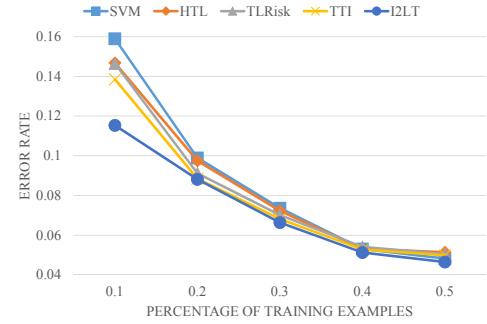


Fig. 4. Average error rate of different algorithms with various percentage of training images from each concept, from 10% to 50% with an increment of 10%.

regularize the intramodal label transfer) are selected from $\{0, 0.5, 1.0, 2.0\}$, $\{0.1, 0.5, 1.0, 2.0\}$ and $\{1.0, 2.0, 5.0, 10.0\}$, respectively. All the parameters are tuned based on a twofold cross-validation procedure on the selected training set, and the parameters with the best performance are selected to train the models.

8.2 Result on Flickr-Wiki Dataset

The first data set is Flickr-Wiki dataset, consisting of a collection of *Flickr* and *Wikipedia* web pages which contains rich media content with images and their text descriptions. We use ten categories to evaluate the effectiveness on the image classification task. To collect text and image collections for experiments, the names of these 10 categories are used as query keywords to retrieve the relevant web pages from *Flickr* and *Wikipedia*. Both web sites return many web pages in response to the submitted queries. Figure 2 illustrates some examples of retrieved images, Table 1 shows the number of occurrence pairs crawled from Flickr by using different query words, and Table 3 shows the number of Wiki articles retrieved from the subcategories of each topmost category. For example, these subcategories contain the breeds of animals (e.g., bird, horse, dog, and cat), and the list of buildings, mountains and waterfalls.

Flickr is an image sharing web site, storing many user-shared images and their textual descriptions in the forms of textual tags and comments. For *Wikipedia*, we have also retrieved the relevant web pages in the subcategories. In each crawled web page, the images and the surrounding

TABLE 4

Comparison of error rate of different algorithms with (a) two training images (b) ten training images. The smallest error rate for each category is in bold.

(a) Two training images

Category	SVM	HTL	TLRisk	TTI	I2LT
birds	0.3293±0.0105	0.3293±0.0124	0.2817±0.0097	0.2738±0.0080	0.2523±0.0042
buildings	0.3272±0.0061	0.3295±0.0041	0.2758±0.0023	0.2329±0.0032	0.1985±0.0023
cars	0.2529±0.0059	0.2759±0.0048	0.2639±0.0032	0.1647±0.0058	0.1326±0.0082
cat	0.3333±0.0071	0.3333±0.0060	0.2480±0.0109	0.2525±0.0083	0.2256±0.0024
dog	0.3694±0.0031	0.3694±0.0087	0.2793±0.0161	0.252±0.0092	0.2415±0.0074
horses	0.25±0.0087	0.3±0.0050	0.2679±0.0069	0.2±0.0015	0.1879±0.0093
mountain	0.3311±0.0016	0.3322±0.0009	0.2817±0.0021	0.2699±0.0004	0.2482±0.0010
plane	0.2667±0.0019	0.225±0.0006	0.2758±0.0006	0.2517±0.0011	0.2044±0.0004
train	0.3333±0.0084	0.3333±0.0068	0.2738±0.0105	0.2099±0.0060	0.1910±0.0014
waterfall	0.2693±0.0009	0.2694±0.0016	0.2659±0.0020	0.257±0.0007	0.2241±0.0010

(b) Ten training images

Category	SVM	HTL	TLRisk	TTI	I2LT
birds	0.2639±0.0012	0.2619±0.0015	0.2546±0.0018	0.252±0.0008	0.2314±0.0012
buildings	0.2856±0.0002	0.2707±0.0021	0.2555±0.0014	0.2303±0.0017	0.2214±0.0014
cars	0.3027±0.0073	0.3065±0.0030	0.2543±0.0029	0.2299±0.0031	0.2025±0.0022
cat	0.2755±0.0043	0.2525±0.0038	0.2553±0.0028	0.2424±0.0026	0.2289±0.0052
dog	0.2252±0.0039	0.2343±0.0037	0.2545±0.0031	0.2162±0.0027	0.2015±0.0024
horses	0.2667±0.0019	0.2500±0.0021	0.2551±0.0016	0.2383±0.0013	0.2192±0.0018
mountain	0.3176±0.0010	0.3097±0.0003	0.2541±0.0011	0.2626±0.0007	0.2312±0.0004
plane	0.2667±0.0009	0.2133±0.0008	0.2546±0.0005	0.2567±0.0012	0.1815±0.0004
train	0.2624±0.0029	0.2716±0.0118	0.2552±0.0025	0.2346±0.0031	0.2123±0.0026
waterfall	0.2611±0.0008	0.2435±0.0009	0.2555±0.0016	0.2546±0.0007	0.2252±0.0008

TABLE 5

The number of topics (i.e., the rank of matrix S) used for learning the transfer function in topic space from 2,000 co-occurrence pairs with two and ten training examples.

Category	Two examples	Ten examples
birds	15	28
buildings	75	96
cars	7	16
cat	11	15
dog	7	14
horses	5	7
mountain	5	9
plane	17	21
train	8	6
waterfall	19	28

text documents are used to learn the alignment between text and images. It is worth noting that these co-occurrence pairs used to align the image and text modalities do *not* contain any labeled images in the training set. In other words, no images in the co-occurrence pairs are labeled, and hence, these pairs are unlabeled. In fact, in our algorithm, we do not need the labels of these pairs to learn label transfer. These unlabeled pairs are only used to model the correlation between the two modalities.

For images, visual features are extracted to describe these images. For the sake of fair comparison, we use the same vocabulary of visual words to represent images as those used by the compared algorithms in previous work [35]. These include the 500 dimensional bag of visual-words (BOVW) based on SIFT descriptors [27]. For the text documents, we normalize the textual words by removal of stop words and stemming, and use their frequencies as textual features. For

each category, the images are manually annotated to collect the ground truth labels for training and evaluation as shown in Table 2. Nearly the same number of background images are collected as the negative examples. These background images do not contain the objects of the categories. It is worth noting that these image categories are not exclusive which means that one image can be annotated by more than one category.

First, in Figure 3 and 4, we report the performances of different algorithms with varying numbers of training images. For each category, the same number of images from the background images are used as the negative examples. Then average error rate is shown to evaluate the performance for image classification tasks. To learn the transfer function, 2,000 co-occurrence pairs are collected to learn the alignment between texts and images for label transfer. Since each image can be assigned more than one label, the error rate is computed in binary-wise fashion.

We note that as shown in Figure 3, a small number of training images is the most interesting case for our algorithm, because it handles the challenging cases when an image category do not have much past labeling information for the classification process. In order to validate this point, in Figure 3, we compare the average error rates over all categories with varying number of auxiliary training examples. It demonstrates the advantages of our methods when there are an extremely small number of training images. This confirms our earlier assertion that our approach can work even in the paucity of auxiliary training examples, by exploring the correspondence between text and images. In Tables 4(a) and 4(b), we compare the error rate of different algorithms for each category with two and ten auxiliary training images respectively. We note that Table 4(a) (a) shows the results with a *extremely smaller number of train-*

TABLE 6
Comparison of Mean Average Precision (MAPs) on NUS-WIDE dataset.

Algorithms	MAPs
SVM	0.2078
HTL	0.2563
TLRisk	0.3048
TTI	0.3418
I2LT	0.4042

ing images, and the proposed scheme outperforms the compared algorithms on all the categories. If we continue to increase the number of training examples to a large enough level, as shown in Figure 4, the advantage achieved by the label transfer algorithm gradually diminishes. This is expected since with sufficiently training examples, there is no need to leverage the cross-modal labels to enhance the classification accuracy.

Also, Table 5 lists the number of topics (i.e., the rank of matrix S) used for learning the transfer function in topic space from 2000 co-occurrence pairs with two and ten training examples. It shows that for most of categories with only a small number of topics, the learned label transfer model works very well. This also provides evidence of the advantages of the parsimony principle in semantic translation. However, this criterion is not absolute or unconditioned, but with the premise that the observed training examples and co-occurrence pairs can be fit by the learned model. For complex categories with many aspects, it often uses more topics to establish the correspondence between the heterogeneous domains. For example, as the appearances of “buildings” vary largely with lots of variants, more topics are needed to explain the correspondence between these variants than the categories with relatively uniform appearances. But as long as the training data can be explained, the models with fewer topics are preferred for the improved generalization performance.

8.3 Result on NUS-WIDE Dataset

The second dataset we use to evaluate the algorithm is NUS-WIDE [12], which is a real-world image dataset that contains 269,648 images downloaded from Flickr. Each image has a number of textual tags and is labeled with one or more image concepts out of 81 concepts. The 186,577 image-text pairs belonging to the 10 largest concepts are selected as co-occurrence pairs. Similar to the above Flickr-Wiki dataset, the images are represented by 4096-D Convolutional Neural Network (CNN) features by AlexNet [21] and the image tags are represented by 1000-D word occurrence feature vectors.

Figure 5 plots the comparison of Average Precision (AP) for 81 concepts on NUS-WIDE dataset. Average Precision measures how well an algorithm ranks the positive examples higher than the negative examples [49]. It is a widely-used metric in comparing between different classification algorithms especially with an imbalanced sets of the positive and negative examples. From the result, we can find that on 67 out of 81 concepts, the proposed I2TL outperforms the other compared algorithms. Table 6 compares the Mean Average Precision (MAPs) over 81 concepts on NUS-WIDE dataset.

TABLE 7
Comparison of CCA-type cross-modal retrieval models and the proposed I2LT algorithm. The performance reported in the table is Precision@20. The results are an average from five random splits of database/query splits, and the standard deviations are on an interval of [0.52% – 1.34%].

Algorithms	I2I	T2I
CCA(V+T)	42.44	42.37
CCA(V+T+K)	48.06	50.46
CCA(V+T+C)	44.03	43.11
I2LT+BOVW	56.34	56.85
I2LT+CNN	68.72	69.20

It is worth noting that the learned intermodal label transfer function measures the cross-modal relevance. It can be used to retrieve the relevant the images given a query of text description, and vice versa. Thus, we test the cross-modal retrieval with the learned transfer function. Specifically, following the experimental setup in [18], we consider two scenarios. (1) The Image to Image (I2I) search, i.e., an image is used as a query to search the relevant images with the same label; (2) The Text to Image (T2I) search, i.e., the input query is a text description and the output is a list of relevant images. We can also perform an Image to Text (I2T) search, where an image is used as input query to search for the relevant text descriptions. However, the I2T result on NUS-WIDE was not reported in [18]. For the sake of a straight comparison, we skip the I2T search in this paper too.

We follow the same evaluation protocol as [18]: from the test set, 1,000 samples are randomly sampled and used as the queries, 1,000 as the validation set, and the remaining ones are retrieved. A retrieved output is considered as relevant to an input query if they have the same label. The experiments are repeated five times, and we report the top-20 precision averaged over the five random database/query splits.

Table 7 compares the retrieval performances by I2LT and the other three CCA variants. Among them, CCA(V+T) refers to the two-view baseline model based on both visual and text features; CCA(V+T+K) refers to the three-view CCA model with visual, text and supervised semantic information; and CCA(V+T+C) refers to the three-view model with unsupervised third view on automatically generated word clusters. More details about these three models can be found in [18]. In testing I2I search, image features are projected into the CCA space and the learned I2LT space (cf. Eq. (6)) respectively, and then we use them to retrieve the most relevant images from the dataset. For the fair comparison with these CCA variants trained with the original BOVW features, we report the retrieval accuracies by I2LT with BOVW features and CNN features in the table.

8.4 Zero-Shot Label Transfer on CUB200 and Oxford Flower-102 Datasets

We used two datasets to test the algorithm for the zero-shot label transfer. The first one is CUB200 Birds dataset [46] which consists of 200 species of birds in 6033 images. The corresponding wikipedia articles are collected by using the name of these birds as query keywords, ending up with

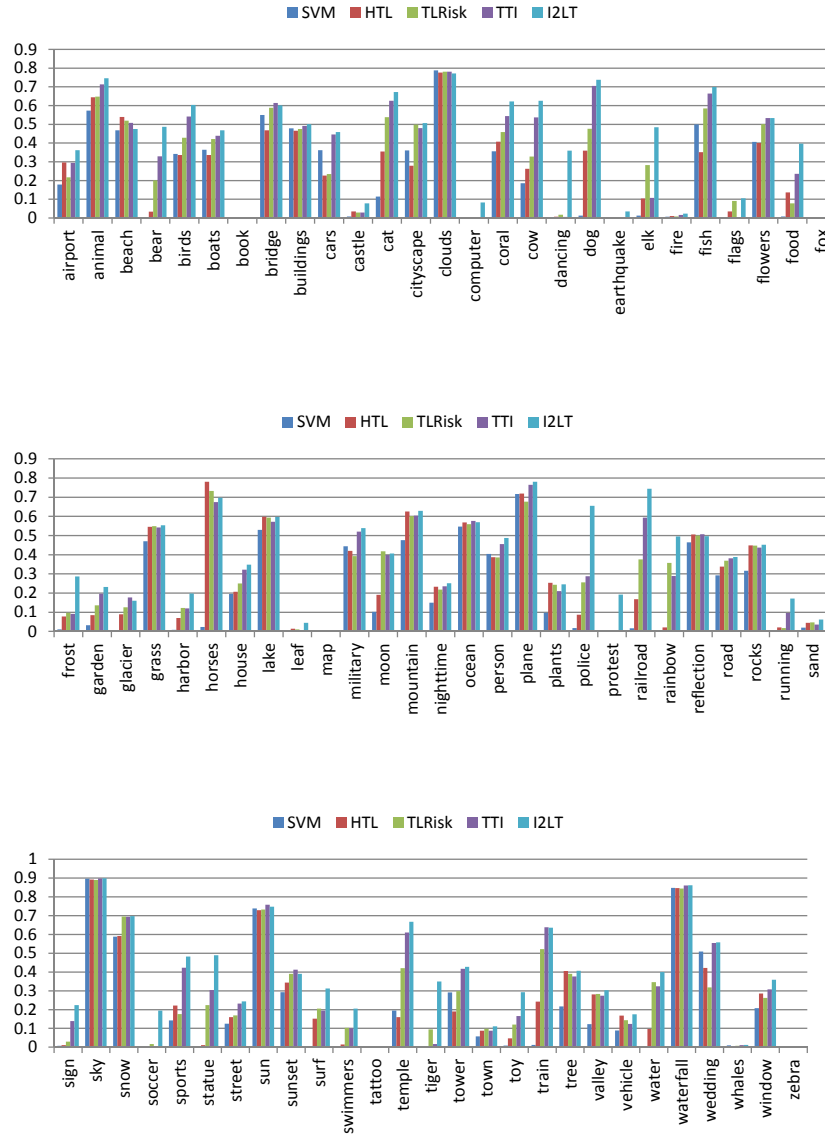


Fig. 5. Comparison of Average Precision (AP) for 81 concepts on NUS-WIDE dataset. The plot is better viewed in color.

200 articles as the text descriptions [16]. The second dataset is Flower102 with 102 classes of flowers in 8189 images [31]. Different from CUB200, the text articles, one for each flower class, are collected not only from Wikipedia, but also from Plant Database, Plant Encyclopedia, as well as BBC articles [16].

Both datasets extracted 2569 dimensional Classme features as an intermediate semantic representation of the input images. For the text modality, TF-IDF (Term-Frequency and Inverse Document Frequency) features are extracted from each article, followed by reducing 8875-dimensional TF-IDF features to 102 dimension with Cluster Latent Semantic Indexing (CLSI) algorithm. The resultant dataset with text descriptions is publicly available [16]².

Five-fold cross validation over the was adopted to test the algorithm, where 4/5 classes were used as seen classes and the other 1/5 of classes as unseen ones. Then the datasets are split into training and test sets according to

the seen and unseen classes, where the images and the corresponding articles of those seen classes constitute the co-occurrence pairs. The five-fold cross-validation over the seen classes is used to decide the hyper-parameters. Following [16], we report the average AUC (Area Under ROC Curve) over five-fold cross-validation to evaluate the performance.

We considered four state-of-the-art zero-shot learning algorithms as baselines, namely (1) Gaussian Process Regressor (GPR) [41], (2) Twin Gaussian Process (TGP) [8], (3) Nonlinear Asymmetric Domain Adaptation (DA) [22], as well as (4) WAC (Write A Classifier) [16].

We report the comparative results on the two datasets in Table 8. We can see that the proposed algorithm outperforms the others in terms of average AUC. The performance improvement is partly attributed to the fact that the proposed approach prefers the concise label transfer model by imposing the trace norm regularizer. This preference plays an important role considering that only very rare positive samples are available for unseen classes in text and image

2. <https://sites.google.com/site/mheltoseiny/computer-vision-projects/zero-shot-learning>

TABLE 8

Comparison of Zero-Shot classifiers on CUB200 and Flower102. Average AUC and its standard deviation are reported.

Algorithms	CUB200	Flowers102
GPR	0.52±0.001	0.54±0.02
TGP	0.61±0.02	0.58±0.02
DA	0.59±0.01	0.62±0.03
WAC	0.62±0.02	0.68±0.01
Our approach	0.67±0.02	0.73±0.01

modalities (There exists no image examples for zero-shot learning!). With extremely rare examples, adopting a concise cross-modal transfer model can minimize the over fitting risk effectively as both modalities have much high dimensionality of feature representations. Actually, the resultant label transfer matrices \mathbf{S} are only of rank 35 ± 3 on CUB200 dataset and of rank 28 ± 5 on Flower102 dataset over five-fold cross-validation.

We also compare the classification accuracies with various types of output embedding models [2] on the extended CUB dataset in Table 9. This dataset extends CUB200 dataset to have 11,788 images from 200 bird species. For a fair comparison, the same zero-shot split as in [1] [2] is used, where 150 classes are used for the training and validation, and the remaining 50 disjoint classes are used for testing. The average per-class accuracy is reported on the test set for each compared algorithm. From the comparison, we can find the proposed algorithm outperforms the compared types of embedding algorithms. This can be attributed to the proposed algorithm which does not only use input and output embeddings to learn the transfer function, but also applies the learned transfer function to combine multiple labels of source texts to annotate the target images. On the contrary, these existing embedding models only output the compatibility between an input-output pair, without exploring the joint use of multiple source labels to predict on a target image.

Here, we wish to make an additional note on the training of the proposed label transfer model in the zero-shot scenario. In the experiment, we enforce that the label transfer matrix \mathbf{S} is shared across seen and unseen classes. This is possible because the transfer matrix is class-independent, since it aims to capture the inter-modal correlation between images and their corresponding articles no matter which classes they belong to. In this sense, in the training phase, each seen class of images in the training set can also be labeled by transferring the corresponding text labels with the shared transfer matrix learned by minimizing such label transfer errors as in Eq. (4). In this way, we fully explore the image labels of seen classes in the training set to learn the shared transfer matrix. This does not violate the zero-shot assumption that no image labels of unseen classes should be involved in the training algorithm. The experiment results also show that, without this training strategy, the proposed approach only achieved an average AUC of 0.64 and 0.70 on CUB200 and Flower102 by learning a separate transfer matrix for each unseen class, justifying this transfer matrix sharing strategy.

TABLE 9

Comparison of Zero-Shot classifiers on the extended CUB dataset. We compare with the Structured Joint Embedding (SJE) on various types of output embeddings. Classification accuracies are reported.

Algorithms	Extended CUB
Word2Vec(φ^W)	28.4%
GloVe(φ^G)	24.2%
Bag-of-Words(φ^B)	22.1%
WordNet(φ^H)	20.6%
human($\varphi^{0,1}$)	37.8%
human(φ^A)	50.1%
Our approach	55.3%

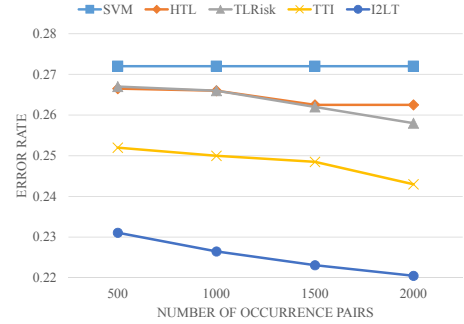


Fig. 6. Average error rate of different algorithms with varying number of image-text pairs to learn the intermodal transfer function.

8.5 Impact of the Size of Co-occurrence Pairs

These above results are obtained by using 2,000 pairs of co-occurred text and images. We know the number of co-occurrence text-image pairs play an important role to align the heterogeneous modalities. Therefore, it is instructive to examine the effect of increasing the pair numbers. In Figure 6, we compare the error rates of different algorithms with varying numbers of text-image pairs. The number of pairs is illustrated on the horizontal axis, whereas the error rate is illustrated on the vertical axis. As we can see, the error rate of the proposed I2LT algorithm decreases with an increasing number of pairs because more information is exploited to align text and image domains. We also note that its improvement is more significant than other algorithms when more text-image pairs are involved. This shows that I2LT is more resistant against the noisy co-occurrence pairs of texts and images by jointly modeling the relevance of training labels between the texts and the images used for label transfer. It also demonstrates the advantage of I2TL over the other algorithms.

8.6 Computational Cost

Finally we compare the computational costs made by different algorithms. All the algorithms are conducted on the same cluster server, equipped with Intel Xeon 2.5 GHz 12-Core CPU, and 128 GB physical memory. Table 10 shows the computing time to train and test with the different models. It is shown that SVM is the fastest model to train since it does not involve any labeled text corpus. TLRisk is the second fastest model to train, and the other three models are trained in the comparable time since all of them spend most of time on constructing intermediate representation to transfer the labels. For test, I2LT uses the longest time

TABLE 10

The training and testing time with 2000 co-occurrence pairs and 10 training examples (in seconds).

Algorithms	Training	Testing
SVM	0.6	8.2
HTL	22.6	24.1
TLRisk	17.2	14.5
TTI	25.3	26.1
I2LT	25.4	29.3

because it has to transfer the labels from the intermodality as well as intramodality. But the longer time is compensated by the more accurate test results as shown above.

9 CONCLUSION

In this paper, we presented a method to jointly transfer labels within and across modalities for an effective image classification model. This method is designed in order to alleviate the dual issues of scarce labels and high semantic gaps which are inherent for the images. The label transfer process is designed with the development of a transfer function, which can convert the labels from text to images effectively. We show that the transfer function can be learned from the co-occurrence pairs of texts and images as well as a small size of training images. We follow the parsimonious principle to develop a common representation to align texts and images with as few topics as possible in the label transfer process. For prediction, we **do not** assume that a test image comes with any text description, and the labels of the text corpus can be propagated to annotate the test image by the learned transfer function. We show superior results of the proposed algorithm for the image classification task as compared with state-of-the-art heterogeneous transfer learning algorithms.

ACKNOWLEDGEMENT

The first author was partly supported by NSF grant 16406218. We also would like to thank the anonymous reviewers for bringing the zero-shot learning problem into our attention, which inspires us to study the applicability of the proposed approach to this problem.

REFERENCES

- [1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 819–826. IEEE, 2013.
- [2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.
- [3] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of International Conference on Machine Learning*, 2007.
- [4] F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of International Conference on Machine Learning*, 2004.
- [5] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.
- [6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, January 2003.
- [7] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.
- [8] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28–52, 2010.
- [9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [10] J.-F. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion, September 2008.
- [11] Y. Chen, T. V. Nguyen, M. Kankanhalli, J. Yuan, S. Yan, and M. Wang. Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(11):1992–2003, 2014.
- [12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR’09)*, Santorini, Greece., July 8–10, 2009.
- [13] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [14] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *Proceedings of Advances in Neural Information Processing Systems*, 2008.
- [15] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1375–1381. IEEE, 2009.
- [16] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2584–2591. IEEE, 2013.
- [17] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multi-modal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):303–316, 2014.
- [18] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2013.
- [19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [20] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, 1999.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105.
- [22] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.
- [23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.
- [24] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.
- [25] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1134–1148, June 2014.
- [26] W. Li, L. Niu, and D. Xu. Exploiting privileged information from web data for image categorization. In *European Conference on Computer Vision*, September 2014.
- [27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [28] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):824–830, 2014.
- [29] S. Moon, S. Kim, and H. Wang. Multimodal transfer deep learning for audio visual recognition. *arXiv preprint arXiv:1412.3121*, 2014.
- [30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.
- [31] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP’08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.

Wei Liu Dr. Wei Liu received the M.Phil. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA in 2012. Currently, he is a research staff member of IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, and holds adjunct faculty positions at Rensselaer Polytechnic Institute and Stevens Institute of Technology. He has been the Josef Raviv Memorial Postdoctoral Fellow at IBM T. J. Watson Research Center for one year since 2012. His research interests include machine learning, data mining, computer vision, pattern recognition, image processing, and information retrieval. Dr. Liu is the recipient of the 2011-2012 Facebook Fellowship and the 2013 Jury Award for best thesis of Department of Electrical Engineering, Columbia University. Dr. Liu has published over 70 papers in peer-reviewed journals and conferences including Proceedings of IEEE, IEEE Transactions on Image Processing, NIPS, ICML, KDD, CVPR, ICCV, ECCV, MICCAI, IJCAI, AAAI, SIGIR, SIGCHI, DCC, etc. His recent papers win CVPR Young Researcher Support Award and Best Paper Travel Award for ISBI 2014.



Charu Aggarwal Dr. Charu Aggarwal is a Distinguished Research Staff Member at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from Massachusetts Institute of Technology in 1996. He has since worked in the field of performance analysis, databases, and data mining. He has published over 135 papers in refereed conferences and journals, and has been granted over 50 patents. He has served on the program committees of most major database/data mining conferences, and served as program vice-chairs of the SIAM Conference on Data Mining, 2007, the IEEE ICDM Conference, 2007, the WWW Conference 2009, and the IEEE ICDM Conference, 2009. He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering Journal from 2004 to 2008. He is an action editor of the Data Mining and Knowledge Discovery Journal, an associate editor of the ACM SIGKDD Explorations, and an associate editor of the Knowledge and Information Systems Journal. He is a fellow of the ACM (2013) and the IEEE (2010) for "contributions to knowledge discovery and data mining techniques".



Thomas S. Huang Prof. Thomas Huang received his Sc.D. from MIT in 1963. He is a full-time faculty with Beckman Institute at University of Illinois at Urbana-Champaign. He was William L. Everitt Distinguished Professor in the Department of Electrical and Computer Engineering and the Coordinated Science Lab (CSL), and was a full-time faculty member in the Beckman Institute Image Formation and Processing and Artificial Intelligence groups. His professional interests are computer vision, image compression

and enhancement, pattern recognition, and multimodal signal processing.